基盤技術研究促進事業

平成１３年度採択

「高信頼・低消費電力サーバの研究開発」
基本契約成果報告書

委託先　富士通株式会社

平成１８年５月

独立行政法人　新エネルギー・産業技術総合開発機構

目　次

まえがき

サーバと呼ばれる大型のコンピュータシステムは、従来から気象シミュレーションなどの科学技術計算や OLTP(Online Transaction Processing)など事務計算分野で広く利用されてきた。これらの場合では、それぞれの分野に適したシステムが使われてきた。たとえば、科学技術計算分野ではベクトル計算機、事務計算分野ではメインフレーム計算機や RISC プロセッサを複数共有メモリで結合したシステムなどである。

一方、インターネットの普及、ブロードバンド時代の到来を見据えて、Web を利用したさまざまな業務が行われるようになってきた。コンサートチケットの販売、オークション、オンライン銀行などがその例である。将来的には、企業での間接業務(人事管理、経理など)も、IDC(Internet Data Center)や ASP(Application Service Provider)と呼ばれる大規模な計算機リソースにアウトソーシングされるようになると予想される。さらに、電子政府の推進により、国民と政府の間の情報交換も、このブロードバンドインターネットをベースに行われるようになると考えられる。

本委託研究はこのような状況に鑑み、ブロードバンドインターネット時代に適したサーバシステムの研究開発を目指すものである。

1.　研究の目的と目標

ブロードバンドインターネット時代のサーバシステムに必要な機能について考察する。インターネットに接続されたサーバは、インターネットからの種々の要求を確実に時間遅延なく処理できることが必須である。インターネットのような全世界に開かれたネットワークでは予めその負荷を予測できないことが多いため、解決策としてサーバの規模を大きくして対応している。(アウトソーシングセンタが構築されるのも、このような大規模サーバシステムを構築できるのはごく一部の大企業のみであるためである。)　しかし、システムが大規模化すればその信頼性は一般に低下する。また、この種のサーバシステムは 24 時間稼動を要求される。したがって、たとえ障害が発生してもシステムはダウンしない、あるいは障害の可能性を未然に察知し、それに自律的に対処できるようなシステムを作りあげることが必須になる。すなわち、大規模サーバシステムの信頼性向上が重要な研究開発項目となる。

また、サーバシステムが大規模化すれば、その設置面積も大きくなる。これに対応するために、ブレードと呼ばれる小さな基盤に CPU、メモリ、ディスクを搭載して、このブレードを高密度に多数、実装する方式が提案されている。しかし、この場合、実装手段のほかに、発熱量が膨大になるという重大な問題が生じる。この熱の問題を解決することが、火急の課題である。従って、本事業では、大規模サーバシステムの低消費電力化を研究開発項目とする。

以上をまとめる。本事業では、ブロードバンドインターネット時代のサーバシステムとして、多数のブレードから構成される大規模サーバを考え、当社が今までにメインフレーム、Unix サーバシステムなどの開発で培ってきた技術を利用、改良、拡張しながら、ブレードサーバシステムの高信頼化、低消費電力化を研究開発のターゲットとする。また、このような高信頼、低消費電力サーバを、オーガニックサーバと名付ける。

2.　委託事業の研究成果

2.1　研究スケジュール表

基本計画書に記載した研究スケジュールを下表に示す。

	平成 13 年度	平成 14 年度	平成 15 年度	平成 16 年度	平成 17 年度
高信頼なブレードサーバシステムの研究開発		仕様検討	試作･評価	改良	実用化版開発
低消費電力ブレードサーバシステムを実現する研究開発(CPU)		試作･評価	2 次試作	改良	実用化版開発
低消費電力ブレードサーバシステムを実現する研究開発(ストレージ)		仕様検討･シミュレーション	試作	評価･改良	実用化版開発
高信頼･高性能なインタコネクトの開発		仕様検討　開発	評価	高機能版	開発

中間評価

研究実績を以下に示す。

	平成 13 年度	平成 14 年度	平成 15 年度	平成 16 年度	平成 17 年度
高信頼なブレードサーバシステムの研究開発	→	仕様検討 →	試作･評価 →	改良 →	実用化版開発 →
低消費電力ブレードサーバシステムを実現する研究開発(CPU)	→	試作･評価 →	2 次試作 →	改良 →	評価 →
低消費電力ブレードサーバシステムを実現する研究開発(ストレージ)	→	仕様検討･シミュレーション →	☆予算削減により中止		
高信頼･高性能なインタコネクトの開発		仕様検討 → 開発 →	評価 →	高機能版 →	開発 →

2.2 内容及び成果

【平成 13 年度】

①高信頼ブレードサーバシステムの仕様検討および試作

ブレードから構成される大規模なシステムの信頼性を向上させる方式として，実行時に動的な並列度を変更する機能の開発に着手し，今年度は，各ノードの負荷に応じて起動中のユーザプロセス間で動的にデータを受け渡す機能を開発し，クラスタ制御ソフトであるSCoreに実装した．また，コモディティ製品を集めたクラスタシステムにおいて，その信頼性を確保する手法として多数決による手法を提案した．本手法は同一の処理を複数のクラスタで実行し計算の中間結果/最終結果を互いに比較する．全クラスタの計算結果で多数決の勝者を決定し，最終結果とすることによって，計算結果の信頼性および，クラスタシステムのハードウェアの信頼性を向上させることができる．

②ブレードサーバシステムの低消費電力化の検討および試作，実験

多数のブレードから成るシステムの省電力化技術として，PCの省電力機能を応用し，アイドル状態のノードは積極的にサスペンドさせて必要なときにリジュームする省電力機能をSCoreに実装して評価を行った．また，ストレージシステムの省電力化に関しては，高性能ディスクと低消費電力の大容量ディスクの階層構成で，RAID(Redundant Arrays of Independent Disks)のレスポンス性能を保持したまま消費電力を抑えることを検討した．80%のデータアクセスが20%の領域に集中する仮定のもとでは，高性能ディスクをキャッシュとして使うことで，容量あたりの消費電力を1/3に抑えることができ，大容量ディスクと高性能ディスクで階層RAIDを構成することでさらに電力消費を抑えられることがわかった．

③ブレードサーバシステム実験プラットフォームの開発

①，②の研究項目を遂行するプラットフォームとして，CPUのダイ温度，ディスクの表面温度や消費電力などの測定，およびブレード間の高速通信用ボードの拡張が可能なCPUブレードを開発した．また，CPUブレードの電源投入などの管理を行なうマネージメントブレードも合わせて開発した．これらを組み合わせ，19インチラック2台に200枚のCPUブレードから構成されるクラスタシステムを構築し，上記①，②の研究を推進した．

【平成 14 年度】

① 高信頼なブレードサーバシステムの研究開発

昨年度購入した200台のブレードから構成されるオーガニックサーバ(ブレードサーバ)を用いて、高信頼化技術とプログラム開発環境の研究開発を行った。具体的には、(1)ブレード制御を行うクラスタ OS の機能の一部の試作、(2)同一の処理を複数のクラスタで実行し計算の中間結果/最終結果を互いに比較する多重実行／多数決方式の研究、(3)運用管理ソフトウェアを統一して扱う統合管理基盤の開発、(4)大量のジョブをクラスタシステムで効率よく実行するオーガニックジョブコントロール技術、(5)ブレードサーバのアーキテクチャを活かしたソフトウェアを開発するための開発環境フレームワークとその一部として動作するコンパイラに組み込む分散並列化コード生成機能の開発を行った。

② 低消費電力ブレードサーバシステムを実現する研究開発

試作したオーガニックサーバ上に、ブレードの負荷状況を監視して負荷が低い状態ではブレードをスリープさせることにより省電力化する方式を開発した。

(1) 省電力

多数のブレードから成るシステムの省電力化技術として、Linux のソフトウェアサスペンド機能を用い、アイドル状態のノードは積極的にサスペンドさせて必要なときにリジュームする省電力システムを開発した。

(2) 動的負荷分散

実行時に動的に並列度を変更する HPC 向け負荷分散機能を開発した。システムの状況にあわせて自律的にアプリケーションのプロセス数を増減する方式を開発し、通信ライブラリである LAM/MPI に実装した。そして、異種アプリケーション(Web サーバ、レイトレーシング、HPC)間での動作検証を行った。

③ 高信頼・高性能なインタコネクトの開発

今年度は、大規模なブレードサーバを構成するための 10 ギガビットイーサネットスイッチチップの設計を終了し、製造を開始した。本スイッチチップは 12 個の 10 ギガビットイーサネットポートを持ち、高バンド幅に加えて非常に低レテンシのために、ブレードサーバやクラスタの構築に非常に適している。また、本スイッチチップを評価するための評価ボードと制御用ファームウエアの試作を行った。更に、来年度に開発を予定している性能強化版の 10 ギガビットイーサネットスイッチチップに搭載を予定している、Enhanced XAUI マクロのテストチップの評価を行い、電気で 20 m程の銅線ケーブル上を伝送することに成功した。この IO マクロを使えば、高価な光モジュールを使わずに筐体間の接続が出来るので、大規模なブレードサーバやクラスタを安価に構築することが可能となる。また、この評価結果を反映して改良した eXAUI マクロとテストチップの設計を終了した。来年度にその評価を行う。

【平成 15 年度】

①ブレードサーバの高信頼化

自律基盤を提供するクラスタ OS に関しては、システムの応答時間や負荷情報などを監視する機構、それらを元に最適な処理性能を得られるように Web サーバ数やアプリケーションサーバ数を自律的に調整する機構、必要性能に応じてシステムシミュレーションによりサーバ数などのハードウェア構成を精度良く導出する機構を開発し、システム試作を行った。運用管理機能に関しては、ブレードサーバの構成を統合管理し、構成変更の自動化を行うブレード統合管理基盤の開発・検証を行った。

② ブレードサーバの低消費電力化

今後の低消費電力・高信頼機能の研究のプラットフォームとなるブレードサーバシステムのアーキテクチャ検討と設計を行い、基本部である CPU ブレード、ディスクブレード、スイッチブレードおよび、これらを搭載するシャーシを試作し、評価を行った。本ブレードサーバの主な特長は、機能別ブレードによるシステム構成の自由度と耐故障性の向上、ホスト CPU とは独立に搭載した制御プロセッサによる負荷状況や消費電力などの動作状況を考慮した自律制御の実現、である。

③ ブレードサーバ用のプログラム開発環境の開発

ジョブ制御に関しては、利用者の利便性向上を目的に、任意のバッチシステムを利用するフレームワークの作成などの機能拡張を行った。自律的負荷分散とデータの分割を可能にするコンパイラに関しては、コンパイラと組み合わせて利用する自律負荷分散ライブラリを開発した。大規模ドキュメントを用いた検索アプリケーションに関しては、性能の向上と実証実験を行い、クラスタＯＳと連携して自律的に各アプリケーションで使用するブレードサーバ数を調節し、応答時間を安定化させるように負荷分散できることを確かめた。

④　高性能・高信頼インタコネクトの開発

10 ギガビットイーサネットスイッチチップの評価結果を反映すると共に、高性能 IO マクロ(Enhanced XAUI マクロ)を搭載して、上期に設計試作した性能強化版の 10 ギガビットイーサネットスイッチチップの試験、評価を行い、10 ギガビットイーサネットの信号を銅線ケーブルにより25m 安定に伝送出来ることを確認した。また、スイッチの機能も設計通りに動作することを確認した。これにより、大規模なブレードサーバを高密度かつ経済的に構成するためのインタコネクト用スイッチチップの開発に世界で初めて成功した。高速インタコネクト機能に関しては、大規模ネットワークの構成制御、障害発生時のリカバリー処理を開発して評価を行った。

【平成 16 年度】

① 高信頼・低消費電力を実現するプラットフォームの研究開発

本年度は昨年度に引き続き、高信頼・低消費電力を実現するプラットフォームとして、機能別ブレードで構成するシステムの開発および実験評価を行った。CPU ブレードに関しては、キャパシタ付き CPU ブレードの試作および高速 CPU ブレードの仕様検討を行った。前者は動作中のブレードの挿抜や状況に応じた電力制御を可能とすることで、メンテナンスの自由度向上や低消費電力化に貢献する技術である。

また、機能別ブレード向けストレージ機能の実装を行った。これは、メモリブレードをソケットで結合して１つの大きなストレージキャッシュに見せることで、高速なストレージ機能を実現する技術である。1 ギガビットイーサネットで相互結合した実験システムで性能を測定したところ、ラ

ンダムリードに対して 7500 iops(30 MB/sec) の性能が得られた。

システム全体の制御機構(オーガニック・システム・ウェア)に関しては、システム立ち上げやメンテナンス時におけるハードウェアの追加・削除をサポートする機構、およびスイッチブレードを動的に変更する機能を開発した。さらに、オーガニック・システム・ウェアの上位機能として、ノード上で動作しているアプリケーションをその動作環境である OS ごと異なる物理ノードへ移動させるマイグレーション機能を利用して、システムの運用管理を行う構成制御機構を開発し、機能別ブレードシステム上でその動作を確認した。

② 高信頼・低消費電力を実現するソフトウェアの研究開発

本年度は、昨年度までに試作した自律機能を実現する各種ソフトウェアの機能強化と利用者ビューでの統合を進め、オーガニックサーバを支える自律ソフトウェアシステムの構築を進めた。

クラスタ OS(Phantom)に関しては、高信頼化機能として、管理ノードに障害が発生してプロセスが停止した場合でも、サービスを継続することが可能となる機構を開発した。この機構はサービスを提供しているノードを管理ノードに切り替えることで実現されており、これにより、省電力性を維持しつつ信頼性を向上させることができる。

オーガニックサーバの有用性検証のために開発している検索アプリケーションに関しては、これまで性能ボトルネックとなっていた検索マスタを並列化し、より柔軟な負荷分散が行えるように改良した。また、多数の関連する検索要求を発生させる並列検索結果解析機能を実装した。

大規模 R&D に関しては、データと計算の自律的な動的再分散を、より柔軟にするためのコンパイラ機能を試作・評価した。特に、プロセッサ(ブレード)を物理と仮想の２階層で捕らえ、プロセッサの仮想的な並べ替えを実現することにより、多次元分散での効果的な動的再分散を実現した。

プログラム利用環境のフレームワークに関しては、昨年度までに開発した開発環境フレームワークをベースに、ブレードサーバ向けのアプリケーションを効果的に実行するための利用環境(ジョブの開発環境)を開発した。また、大量ジョブを効率的に実行するための利用技術(ジョブ制御機能)との低レベルな連携をさらに発展させ、実用レベルの連携とするための改善を施した。

ジョブ制御機能に関しては、新たに、並列ジョブとして MPI を利用したプログラムをジョブ投入するための仕組み(MPI エージェント)を開発した。MPI エージェントは実行に必要なノードの確保・ジョブ投入・監視を行うことで、ジョブ制御機能からはあたかも一つのジョブであるかのように振舞う。また、一度の実行で複数のバッチシステムにジョブを投げ分けるバッチドライバ機能と、実行途中に不要なジョブを削除するジョブ打ち切り機能を実装した。さらに、実用レベルのツールとするために、各国語での出力メッセージ対応と構文解析の見直しを行った。

システム管理ソフトウェアに関しては、ブレードシステムのサーバプール構成管理機能、サーバプールからスペアサーバをサービスに追加する機能、およびサーバ故障を検知し、故障したサーバからスペアサーバにブートディスクを自動的に切替え、サービスを継続する自動リカバリ機能の試作・評価を実施した。

③ 高性能・高信頼インタコネクトの開発

大規模なブレードサーバを構成するための 10 ギガビットイーサネットスイッチチップの開発を昨年度に引き続き推進した。今年度は昨年度開発したチップをベースに、高信頼化等の大幅な機能追加と省電力化を図った改良版のスイッチチップの設計・製造・試験を行った。評価の結果、追加機能が設計どおりに動作することと消費電力が 30%以上削減されていることを確認した。さらに、次世代チップの設計に着手し、従来に比べ 2 倍近く高性能なアーキテクチャの基本設計と、10 ギ

ガビット・シリアル伝送を可能とする高性能IO回路のテストチップ設計を行った。

また、広帯域インタコネクトによる高速通信を実現するため、TCP/IPのプロトコル処理の全部あるいは一部をホストCPUからNIC上にオフロードするTOE（TCP Offload Engine）の実現方式を検討し、メッセージのコピーを削減する方式をFPGA上に試作実装した。評価の結果、メッセージ送受信の基本機能が動作していること、従来に比べCPU使用率が減少し、通信スループットが向上していることを確認した。

【平成17年度】

① 高信頼・低消費電力を実現するプラットフォームの研究開発

本年度は、高信頼・低消費電力を実現するプラットフォームとして、昨年度に開発した機能別ブレードの改良を行なった。具体的には、CPUブレードに関しては搭載するキャパシタの大容量化と制御機構により、ブレードサーバの低消費電力化技術を強化した。スイッチブレードに関しては、必要帯域に応じた使用ポート数制御を行うことにより、消費電力削減を実現した。また、機能別ブレード向けストレージ機能の高信頼化に関して、高速性と信頼性を両立させるため、書き込みにキャッシュを使わずログ形式でシーケンシャルに書き込む方式を実装した。

システム全体の制御機構(オーガニック・システム・ウェア)に関しては、一部のシャーシに負荷が集中し、電力や熱がシャーシの許容量を越える可能性があるとき、キャパシタ付きCPUブレードの電力の利用や、周波数制御を行いながら別のシャーシに実行ノードを移動させる(マイグレーション)ことにより、電力や熱が許容範囲内で動作する自律制御機構の開発と評価を実施した。

② 高信頼・低消費電力を実現するソフトウェアの研究開発

本年度は、昨年度に引き続き、各種ソフトウェアの機能強化と利用者ビューでの統合を進め、オーガニックサーバを支える自律ソフトウェアシステムの構築を進めた。

クラスタOS(Phantom)に関しては、ノードスケジューリング機構によってファイアウォールの数を増減する機構を開発した。これは運用系のネットワークと管理系のネットワークを分離することによって実現されている。また多重化実行/多数決機能との連携をはかり、自律運用システムとしての適用範囲を広げた。

大規模R&D対応に関しては、外乱負荷の変動によるプロセス間の能力バランスの変化を動的に評価する機能と、その評価値に応じてプロセスの擬似的休止・再開を含む負荷バランス調整を行う機能を開発し、自律的アプリケーションを生成するコンパイラを実現し、ブレードサーバ上でのシミュレーションにより効果を確認した。

プログラム利用環境のフレームワークに関しては、昨年度までに開発したブレードサーバ向けのアプリケーションを効果的に実行するための利用環境(ジョブの開発環境)をベースに、大量ジョブを効率的に実行するための利用技術(ジョブ制御機能) を含め、性能面や拡張性を考慮した内部情報のDB化、運用の信頼性を考慮した運用監視機能の改善等、適用可能性を拡大するための開発を実施した。

ジョブ制御機能に関しては、昨年度までに開発したジョブ制御機能を基に、利用容易性を向上させるため、JSDL標準化への対応、実行可能スクリプトのためのシェルコマンド(ojcsh)の開発、開発環境フレームワークとの連携機能の開発、およびアボートジョブや手動undoジョブを含む待ち合わせ機能を開発した。

システムの運用管理機能に関しては、昨年度に実装したサーバプール構成管理機能を利用し、

サーバリソースを制御するインターフェース、及びサーバリソース追加時にアプリケーション、ミドルウェアの資源、動作環境の設定処理を自動実行するインターフェースを実装した。これにより、急激な負荷変動が発生した場合でも従来よりも短時間でシステム負荷を平準化できることを確認した。

アプリケーションとその動作環境である OS を短時間で他のノードに移動する瞬時マイグレーション機構において、VLAN を利用した複数ユーザ対応の改良と、瞬時マイグレーション管理機構の高可用化のための監視機構を開発し、機能別ブレードシステム上でその動作を確認した。

③ 高性能・高信頼インタコネクトの開発

昨年度に開発した、省電力化および信頼性とセキュリティ向上のための機能追加を行った 10 ギガビットイーサネットスイッチに対してシステム面からの評価と改良を行った。これにより大規模サーバの高信頼化、省電力化を可能とするインタコネクト用 10 ギガビットイーサネットスイッチの実用化に成功した。引続いて、より高性能、高密度、高信頼な大規模サーバの構築を可能とする、次世代の 10 ギガビットイーサネットスイッチチップ、および 10 ギガビット信号のシリアル伝送を可能とする高性能 IO マクロの全ての論理設計、回路設計、物理設計を行い、本スイッチチップおよび高性能 IO マクロの実現性を確認した。具体的には、昨年度開発の高機能化スイッチに比べて、１チップに従来の 2 倍近いポート(20 ポート)と 5 倍のバッファ容量(2.9 メガバイト)を収容しながら、より短い遅延時間、より少ない消費電力を実現すると共に、10 ギガビット信号のシリアル伝送を可能とする高性能 IO マクロを内蔵した次世代 10 ギガビットイーサネットスイッチの全ての設計作業と実現性の確認を行った。

また、広帯域インタコネクトによる高速通信を実現するため、TCP/IP のプロトコル処理の全部あるいは一部をホスト CPU から NIC(Network Interface Card)上にオフロードする TOE(TCP/IP Offload Engine) の開発を昨年度に引き続き実施した。今年度は昨年度開発した TCP/IP オフロードエンジンの改良を検討し、ホストインタフェースの強化および高速メモリへの対応を行ったエンジン部を開発した。また、通信効率を上げるため、大きなパケットの分割やまとめを行う回路の追加実装も実施した。

2.3 事業終了時における今後の課題

本委託研究の成果については、一部は既に事業化されているが、今後の課題として、以下が挙げられる。

(a) 高信頼・低消費電力を実現するプラットフォームの研究開発

これまでに開発した機能別ブレードシステムのハードウェア、システム全体を制御するファームウェアを活用し、実際の運用形態に即した実験データを集積し、効果を明確化する。

(b) 高信頼・低消費電力を実現するソフトウェアの研究開発

これまでに開発した種々の技術（負荷分散機能、多重化実行、多数決機能）、および、大規模な並列ソフトウェアの開発環境、実行環境、運用管理に関して、実際の運用形態に即した実験データを集積し、効果を明確化する。

(c) 高性能・高信頼インタコネクトの開発

大規模なブレードサーバを構成するための 10 ギガビットイーサネットスイッチチップの開発・評価を継続し、10 ギガビットイーサネットを利用したシステムの普及に備え、10 ギガビットイーサネットスイッチの実用化に向けた研究開発を推進する。

3. 委託事業終了後の事業化計画について

3.1 事業化成果物

これまでに開発した成果を元に、下記の製品化を進めており、一部は既に事業化されている。

(A) ブレードサーバ向け高信頼運用管理機構

システム管理ソフトに、本研究の成果である「ブレードサーバ向け高信頼運用管理機構」を組み込み、製品化する。本管理ソフトにより、複数台のブレードサーバに対して、リソースの物理構成を仮想化し、ブレードサーバに対する OS イメージの作成、メンテナンス資源の配付、VLAN 設定などを自動化することで、必要な時に、サーバに対して最新の業務環境を迅速に構築することを可能とする。また、リソースとサービスの関係を可視化することで、トラブルの発生箇所や影響を把握できるようにし、トラブル発生時の迅速な対応を可能とする。さらに、物理単位だけでなく、同一サービス、同一ソフトウェアなどのグループ単位でサーバの電源制御やメンテナンス資源の配付を行うことができるため、保守作業が簡易化され、保守作業が業務にあたえる影響を少なくし、ブレードサーバの高信頼な運用を可能とする。

(B) オーガニックジョブコントローラ(OJC)

OJC は、ブレードサーバ向け並列アプリケーション開発環境（並列フレームワーク)における動的ジョブ実行制御技術を、グリッド環境の上位層に適用した製品である。OJC を用いることにより、大量のジョブの投入､ジョブのワークフロー的実行およびそれらの管理を GUI ベースの操作で効率的に行うことができる。OJC では、投入する大量のジョブに名前を付けて構造的に管理することができるため､ジョブの一括削除や部分削除､部分的なジョブの再投入などを簡単に行うことができる｡また、ジョブの待ち合わせの記述ができるため､特定のジョブの実行状況に関連付けて別のジョブを動的に投入することもできる。この OJC により、パラメタサーチ、パラメタスウィープを含む、簡易的なワークフロー記述を行うことができる。利用者は、複数の各種計算機ノードで実行させるための実行可能プログラム、必要に応じて入力データ、およびこれらのプログラムをどのように実行するかを記述した OJC スクリプトを用意するだけで､大量の計算シミュレーションの繰り返し実行などの作業において､いろいろな局面で介在する人手による作業や無駄なジョブ投入を排除することで､全体の作業コストを短縮することが可能となる。

(C) HPF(High-Performance Fortran)コンパイラ

HPF は、ブレードサーバなどの分散メモリ型環境での科学技術計算ではデファクト標準と言える並列プログラミング言語である。日本では HPF 推進協議会(http://www.hpfpc.org/)がその普及・推進の活動を続けている。

分散メモリ型環境での計算では、現在は並列言語よりも、Fortran や C のプログラム中で MPI(Message Passing Interface)ライブラリを呼び出す方法が主に使われている。MPI は並列計算機から高い性能を引き出すが、プログラミングの生産性が低いことが問題とされている。

本製品（予定）はプログラミングの生産性と実行性能を高いレベルで両立させる HPF 言語コンパイラである。生成コードは MPI ライブラリを直接使用して通信を行うという特徴を持つので、MPI プログラムを生成するツールとして使用することも可能である。HPF 推進協議会が推奨する HPF/JA 仕様をサポートするため､ループの自動並列化を使った楽なプログラミングから､通信を効率化するきめ細かな指示まで､利用者の必要レベルに応じた記述が可能である｡委託研究成果が生かされたコード生成技術により、HPF の提供する複雑なデータ分散種別に対しても高い性能が得られる。

(D) 10 ギガビットイーサネットスイッチチップ

世界初の 10 ギガビットイーサネット 12 ポートの 1 チップスイッチであり、次のような特長を有している。

・レイヤ 2 でのスイッチングを基本機能にするとともに、高性能のスイッチコアとメモリ制御方式を開発することにより、スイッチング処理に必要な高速バッファメモリや高速 I/O マクロを含め、12 個の 10 ギガビットイーサネットポートを 1 チップ（チップサイズ：256 平方ミリメートル）に集積した。

・チップ上の複数のメモリブロックと、それらを結合する相互結合網を効率良く利用して、高速大容量で多ポートの共有メモリをチップ上に実現する新たな方式(Multi-port Stream Memory)により、10 ギガビットイーサネットポート 12 個が、読み出しと書きこみの 2 つの動作を同時に実行できる 240 ギガビット毎秒の高いバンド幅を実現した。

・到着したパケットを短い遅延時間で出力側に送るために、共有メモリの新しいスケジューリング制御方式を開発し、従来、数マイクロ秒以上かかったスイッチの遅延時間を、450 ナノ秒と従来の 4 分の 1 以下にした。

3. 2 想定市場と委託先における事業の位置づけ

当社の事業領域である、企業ユーザ向けの IT システムを構成する製品として位置づけている。

3.3 事業化体制図

下記に現在想定される事業主体を示す。

(A) ブレードサーバ向け高信頼運用管理機構

富士通株式会社 プラットフォーム技術開発本部

(B) オーガニックジョブコントローラ

富士通株式会社 ソフトウェア事業本部

(C) HPF コンパイラ

富士通株式会社 ソフトウェア事業本部

(D) 10 ギガビットイーサネットスイッチチップ

富士通株式会社 電子デバイス事業本部

3.4 事業化シナリオとスケジュール

下記のスケジュールで事業化を予定している

(A) ブレードサーバ向け高信頼運用管理機構

2004 年 2 月に製品化済。

(B) オーガニックジョブコントローラ

2004 年 8 月に製品化済。

(C) HPF コンパイラ

2007 年 4 月以降に製品化予定。

(D) 10 ギガビットイーサネットスイッチチップ

2005 年 10 月に製品化済。

4. 産業財産権の取得状況

本委託研究の成果として出願した特許の年度毎の出願件数を下表に示す。

	平成 14 年度	平成 15 年度	平成 16 年度	平成 17 年度	合計
特許出願件数	9	9	25	10	53

5. 外部発表等の状況

下表に年度毎の発表件数を示し、次節以降に各発表の概要を示す。

発表項目	平成 13 年度	平成 14 年度	平成 15 年度	平成 16 年度	平成 17 年度	合計
1. 論文発表数	0	0	1	0	1	2
2. 学会発表数	1	0	7	1	5	14
3. ﾌﾟﾚｽ発表数	0	7	7	1	2	17
4. その他	0	1	1	2	0	4

5.1 論文発表

(a) 登録研究員のみの場合

1	論文名	情報処理学会論文誌：コンピューティングシステム(ACS)
	巻、ページ、年	Vol.44、pp.89-100、2003-08
	著者名	松原 正純、鈴木 和宏、勝野 昭
	所属	富士通株式会社
	題名	自律コンピューティングに向けた HPC 向け動的負荷分散機構
2	論文名	情報処理学会論文誌：コンピューティングシステム(ACS)
	巻、ページ、年	Vol.46 No.SIG 12(ACS 11)、pp.245-254、2005-08
	著者名	鈴木 和宏、松原 正純、勝野 昭
	所属	富士通株式会社
	題名	自律運用システム Phantom における高可用化方式の実装

5.2 学会発表

(a) 登録研究員のみの場合

3	学会名	「ハイパフォーマンスコンピューティングとアーキテクチャの評価」に関するワークショップ(HOKKE-2002)
	発表日	2002 年 3 月 7 日
	発表者名	鈴木 和宏、勝野 昭、木村 康則
	発表者所属	富士通株式会社(論文上は株式会社富士通研究所)
	題名	HPC 向け大規模クラスタシステムにおける省電力機能の実装

4	学会名 発表日 発表者名 発表者所属 題名	先進的計算基盤システムシンポジウム SACSIS2003 2003年5月30日 松原 正純、鈴木 和宏、勝野 昭 富士通株式会社 自律コンピューティングに向けたHPC向け動的負荷分散機構
5	学会名 発表日 発表者名 発表者所属 題名	並列/分散/協調処理に関するサマーワークショップ(SWoPP 2003) 2003年8月5日 鈴木 和宏、松原 正純、勝野 昭 富士通株式会社 クラスタ計算機システムにおけるノード仮想化方式
6	学会名 発表日 発表者名 発表者所属 題名	Server Blade Summit 2004月3月10日 中川 幸洋 Fujitsu Laboratories of America, Inc. A Single-Chip 10Gbps Ethernet Switch for New Generation of Server,Storage Applications.
7	学会名 発表日 発表者名 発表者所属 題名	ECMWF(European Centre for Medium-Range Weather Forecasts) High-Performance Networking Workshop in UK 2004年9月23日 中川 幸洋 Fujitsu Laboratories of America, Inc. A Single-Chip 10Gbps Ethernet Switch and Its Switch Box XG800
8	学会名 発表日 発表者名 発表者所属 題名	先進的計算基盤システムシンポジウム SACSIS2005 2005年5月18日 鈴木 和宏、松原 正純、勝野 昭 富士通株式会社 自律運用システムPhantomにおける高可用化方式の実装
9	学会名 発表日 発表者名 発表者所属 題名	SWoPP 武雄 2005 並列/分散/協調処理に関する『武雄』サマー・ワークショップ 2005年8月4日 小林 伸治、安島 雄一郎、新家 正総 富士通株式会社 NICメモリを用いたZero Copy Socket方式の実装評価

(b) 登録研究員以外を含む場合

10	学会名 発表日 発表者名 発表者所属 題名	PSE (Problem Solving Environment) ワークショップ 2003 年 7 月 上田 晴康*1、吉田 武俊*2、安里 彰*2 *1 富士通株式会社、*2 株式会社富士通研究所 ジョブ投入と待ち合わせを記述できるスクリプト言語 オーガニックジョブコントローラ - CAD-Grid への適用 -
11	学会名 発表日 発表者名 発表者所属 題名	並列/分散/協調処理に関するサマーワークショップ(SWoPP 2003) 2003 年 8 月 上田 晴康*1、吉田 武俊*2、安里 彰*2 *1 富士通株式会社、*2 株式会社富士通研究所 ジョブ投入と待ち合わせの出来るジョブ制御スクリプト： オーガニックジョブコントローラの試作
12	学会名 発表日 発表者名 発表者所属 題名	Hot Chips 15 - A Symposium on High-Performance Chips 2003 年 8 月 19 日 Takeshi Shimizu, Yukihiro Nakagawa, Sridhar Pathi, Yasushi Umezawa, Takashi Miyoshi, Yoichi Koyanagi, Takeshi Horie, Akira Hattori Fujitsu Laboratories of America, Inc. A Single Chip Shared Memory Switch with Twelve 10Gb Ethernet Ports
13	学会名 発表日 発表者名 発表者所属 題名	日本コンピュータ化学会 2003 秋季年会 2003 年 10 月 稲田 由江*2、小野寺 聡*1、安里 彰*2、松浦 東*2 *1 富士通株式会社、*2 株式会社富士通研究所 半経験的分子軌道計算プログラム MOS-F の PC クラスタ向け並列化
14	学会名 発表日 発表者名 発表者所属 題名	2005 Symposia on VLSI Technology and Circuits 2005 年 6 月 17 日 Yasuo Hidaka,Weixin Gai,Hideki Osone,Yoichi Koyanagi, Jian Hong Jiang,Takeshi Horie Fujitsu Laboratories of America, Inc. Gain-Phase Co-Equalization for Widely-Used High-Speed Cables
15	学会名 発表日 発表者名 発表者所属 題名	The sixth International Symposium on High Performance Computing (ISHPC-VI) 2005 年 9 月 7 日 岩下英俊、青木正樹 富士通株式会社 Mapping Normalization Technique on the HPF Compiler fhpf

16	学会名 発表日 発表者名 発表者所属 題名	第 13 回「ハイパフォーマンスコンピューティングとアーキテクチャの評価」に関する北海道ワークショップ(HOKKE-2006) 2006 年 2 月 27 日 岩下英俊*1、岡部寿男*2、杉崎由典*1、青木正樹*1 *1 富士通株式会社、*2 京都大学学術情報メディアセンター LINPACK と FFT による HPF コンパイラ fhpf の生産性の評価

5.3 プレス発表

17	掲載されたマスコミ名 発表日 発表題目	日経 BIZTECH 2002 年 7 月 31 日（掲載日） サーバーの動的負荷分散技術を開発
18	掲載されたマスコミ名 発表日 発表題目	日経インターネットテクノロジ 2002 年 9 月（掲載日） サーバー間で自律的に負荷を分散
19	掲載されたマスコミ名 発表日 発表題目	日本工業新聞 2002 年 12 月 17 日（掲載日） 富士通が新 IA サーバー ブレード 200 台実装
20	掲載されたマスコミ名 発表日 発表題目	日経オープンシステム 2003 年 1 月（掲載日） 自律コンピューティングの理想と現実
21	掲載されたマスコミ名 発表日 発表題目	日経産業新聞 2003 年 1 月 31 日（掲載日） 止まらぬサーバー
22	掲載されたマスコミ名 発表日 発表題目	宝島社 ウルトラ ONE 2003 年 2 月 10 日（掲載日） 自律コンピューティングへの道
23	掲載されたマスコミ名 発表日 発表題目	日経コンピュータ 2003 年 2 月 10 日号（掲載日） 富士通が「自律」技術の製品計画を具体化
24	掲載されたマスコミ名 発表日 発表題目	電波新聞 2003 年 5 月 15 日（掲載日） 12 ポートの 10G ビットイーサネットスイッチ 富士通が 1 チップ化(世界初)
25	掲載されたマスコミ名 発表日 発表題目	化学工業日報 2003 年 5 月 15 日（掲載日） 10 ギガ高速イーサネットスイッチ 富士通が 1 チップ化

26	掲載されたマスコミ名 発表日 発表題目	日刊工業新聞 2003 年 5 月 15 日（掲載日） 10 ギガイーサネットスイッチ 1 チップ化に成功
27	掲載されたマスコミ名 発表日 発表題目	日本工業新聞 2003 年 5 月 15 日（掲載日） 毎秒 10 ギガビットイーサネットスイッチ 1 チップで 12 ポート
28	掲載されたマスコミ名 発表日 発表題目	日経エレクトロニクス 2003 年 6 月 9 日号（掲載日） 12 ポートのスイッチ機能を 1 チップ化
29	掲載されたマスコミ名 発表日 発表題目	FIND (Fujitsu Electronic Device News) 2003 年 11 月 Single-Chip 10G-bit Ethernet Switch MB87Q3050
30	掲載されたマスコミ名 発表日 発表題目	http://www.fujitsu.com/us/news/pr/fma_20040324.html (富士通米国向けホームページ) 2004 年 3 月 24 日 Fujitsu Microelectronics America, Fujitsu Laboratories of America Introduce New 10Gbps Ethernet Switch Chip Featuring 10GBASE-CX4 Interface Support
31	掲載されたマスコミ名 発表日 発表題目	http://www.fujitsu.com/us/news/pr/fla_20041108-01.html (富士通米国向けホームページ) 2004 年 11 月 8 日 Fujitsu Introduces Revolutionary 10Gb Ethernet Switches in the North American Market
32	掲載されたマスコミ名 発表日 発表題目	http://www.fujitsu.com/us/news/pr/fma_20050525.html (富士通米国向けホームページ) 2005 年 5 月 25 日 Fujitsu Announces the Latest Version of 10Gbps Ethernet Switch Chip
33	掲載されたマスコミ名 発表日 発表題目	http://www.fujitsu.com/us/fcpa/news/pr/20051114-01.html (Fujitsu Computer Products of America, Inc. のホームページ) 2005 年 11 月 14 日 Fujitsu Advances Industry Leading 10Gb Ethernet Switches with Link Aggregation, IGMP Snooping & Port Security

5.4 その他

34	種類 巻、ページ、年 著者名 所属 題名	テレビでの放映 2003 年 3 月 13 日 - 富士通株式会社 NHK BS 放送の「経済最前線」のコーナーでオーガニックサーバが紹介された。
35	種類 巻、ページ、年 著者名 所属 題名	雑誌 FUJITSU Vol.54、No.4、pp.298-304、2003-07 西川 克彦*1、服部 彰*2、勝野 昭*1 *1 富士通株式会社(論文上は株式会社富士通研究所)、 *2 Fujitsu Laboratories of America, Inc. オーガニックサーバ
36	種類 巻、ページ、年 著者名 所属 題名	雑誌 FUJITSU Vol.55、No 6、pp.553-558、2004-09 堀江 健志、清水 剛、服部 彰 Fujitsu Laboratories of America, Inc. Fujitsu Introduces Revolutuinary10Gb Ethernet Switches in the North American Market
37	種類 巻、ページ、年 著者名 所属 題名	学位論文 2005 年 2 月 3 日 岩下 英俊 富士通株式会社 科学技術計算を支援する言語処理系に関する研究 - 科学技術計算のためのプログラミングインターフェースとその処理系に関する研究 -

6. 要約

6.1 和文要約

(a) 高信頼なBladeサーバシステムの研究開発

ブレードから構成される大規模なシステムの信頼性を向上させる方式として、実行時に動的な並列度を変更する機能の研究開発を行い、システムの動作状況に応じて自律的にアプリケーションのプロセス数を増減する方式を開発した。また、同一の処理を複数のクラスタで実行し計算の中間結果/最終結果を互いに比較する多重実行/多数決方式の研究、仮想計算機(VM)を用いた動的なシステム制御により高信頼化を図る技術の開発、ブレード制御を行うクラスタOSの仮想化機能による高信頼化の研究、ブレードの仮想化機能を統合管理することによる高信頼運用管理方式の開発なども実施した。これらの研究項目を遂行するプラットフォームとして、CPUのダイ温度、ディスクの表面温度や消費電力などの測定、およびインタフェースボードの拡張が可能なブレードサーバ(オーガニックサーバ)を開発した。

本システムの有効性を評価し高めるために、ブレードサーバとSMPとの性能評価、ブレードサーバ用アプリケーション開発環境と実用アプリケーションの研究開発も実施した。ブレードサーバのアーキテクチャを活かしたソフトウェアを開発するための開発環境フレームワークとその一部として動作するコンパイラに組み込む分散並列化コード生成機能の開発を行った。また、ツールや各種機能の有効性を検証するため、大量のジョブをクラスタシステムで効率よく実行する動的ジョブ実行制御技術(オーガニックジョブコントローラ)を開発した。

さらに、より信頼性を向上させるための技術として、CPU、ディスクなどの機能毎に特化したブレード(機能別ブレード)を開発した。機能別ブレードを活用し、処理を行うCPUブレードとデータを蓄積するディスクブレードを分離し、CPUブレードの故障予見時に、その実行内容を他のCPUブレードに移動させる、瞬間マイグレーション技術を開発した。瞬間マイグレーションでは、まず実行に最低限必要なメモリ内容を移動させて、残りのメモリ内容は処理と並行して移動させるので、利用者からは、数秒間だけしか処理が停止したように見えない。これにより、目標とした、99.999%(年間の停止時間が約5分15秒)以上の稼働率を達成した。

(b) 低消費電力Bladeサーバシステムを実現する研究開発

多数のブレードから成るシステムの省電力化技術として、Linuxのソフトウェアサスペンド機能を用い、アイドル状態のノードは積極的にサスペンドさせて必要なときにリジュームする省電力システムを開発した。この技術により、条件によっては従来システムの1/3に消費電力を低減できる。またピーク電力を抑制する技術として、瞬時に充放電が行えるキャパシタを搭載したブレードを開発した。システムへの負荷が増大した場合、一般にはCPUブレードの消費電力が増大するが、キャパシタに蓄積した電力を必要に応じてCPUブレードに供給することにより、システムに供給される電力を増大させずに、増大した負荷を処理できるブレードサーバを構築することができるようになった。

ストレージシステムの省電力化に関しては,高性能ディスクと低消費電力の大容量ディスクの階層構成で、RAID(Redundant Arrays of Independent Disks)のレスポンス性能を保持したまま消費電力を抑えることを検討した。80%のデータアクセスが20%の領域に集中する仮定のもとでは、高性能ディスクをキャッシュとして使うことで、容量あたりの消費電力を1/3に抑えることがで

き、大容量ディスクと高性能ディスクで階層RAIDを構成することでさらに電力消費を抑えることができることがわかった。しかし、平成15年度の予算削減により、それ以降のストレージの階層化による低消費電力技術の研究開発は中止した。

(c) 高信頼・高性能なインタコネクトの開発

大規模なブレードサーバを構成するための 10 ギガビットイーサネットスイッチチップの第一次版の設計・評価を終了した。本スイッチチップは 12 個の 10 ギガビットイーサネットポートを持ち、高バンド幅に加えて非常に低レイテンシーのために、ブレードサーバやクラスタの構築に非常に適している。その後、この第一次版に対して性能を強化した第二次版の 10 ギガビットイーサネットスイッチチップの設計・評価を行った。この性能強化版チップには、電気信号で 25m程度の銅線ケーブル上を伝送することができる、Enhanced XAUI(eXAUI)マクロを開発して搭載した。このチップを使えば、高価な光モジュールを使わずに筐体間の接続が出来るので、大規模なブレードサーバやクラスタを安価に構築することが可能となる。さらに、障害伝達・処理機能を持つ第三次版 10 ギガビットイーサネットスイッチチップの設計と実現性の確認を行った。

5.2 Abstract

(a) Development of high reliable blade server system

Research and development of changing parallel degree dynamically at the execution time as a method to raise the reliability of the large-scale system which consists of blades was done, and the system which fluctuates the number of processes of application autonomously according to a system status was developed. Moreover, research of the multiplex execution and majority system which performs the same processing by two or more clusters, and compares mutually the middle results and the last result of calculation, development of the technology which realizes high reliability by the dynamic system control using a virtual machine, research to realize high reliability by the virtualization function of the cluster OS which performs blade control, and development of the virtualization function of each blades , were carried out. As a platform which carries out these research items, the blade server (organic server) in which measurement of the die temperature of CPU, the surface temperature of a disk, power consumption, etc. and extension of an interface boards are possible was developed.

In order to evaluate and raise the validity of this system, the performance evaluation between a blade server and a SMP server, and research and development of the application development environment for blade servers and practical application were also carried out. The development environmental framework for developing the software which uses the architecture of a blade server effectively, and the distributed parallel code generation function which is put in a compiler that is a part of the development environmental framework were developed. Moreover, in order to verify a tool and the validity of the various functions, the dynamic job execution control technology (organic job controller) of performing a lot of jobs efficiently with a cluster system was developed.

Furthermore, the functional blade, which specialized for each functions, such as CPU and a disk, was developed as technology for raising reliability more. Using the functional blades, the CPU blade which runs some programs, and the disk blade which stores data were separated, and the instant migration technology which moves the contents of one node to other node at the time of failure foreknowledge of a blade was developed. By the instant migration, since the contents of a memory indispensable for execution are moved first and the remaining contents of a memory are moved in the background, it seems to a user that the stop of execution is only for several seconds. This attained the availability more than 99.999% (annual stop time is about 5 minutes and 15 seconds) that is the goal of this research theme.

(b) Development of low power consumption blade server system

As power-saving technology of the system which consists of many blades, the power-saving system in which the node of an idol state is made to suspend positively and the resume is performed when required using the software suspension function of Linux was developed. With this technology, power consumption can be reduced to one third of conventional systems depending on conditions. Moreover, as peak power control technology, the blade with the

capacitor which can perform quick charge and discharge was developed. Although the power consumption of a CPU blade generally increased when the load to a system increased, the blade server which can process the increased load could be built by supplying the electric power accumulated in the capacitor to a CPU blade if needed, without increasing the electric power supplied to a system.

About power-saving of a storage system, method to hold response time without performance degradation of RAID (Redundant Arrays of Independent Disks) by the layered confguration of a high performance disks and the high-capacity disk of low power were examined. Under assumption which 80% of data access concentrates on 20% of area, it turned out that the power consumption per capacity can be decreased to one third by using a high performance disk as cash. However, the budget reduction in the Heisei 15 fiscal year stopped research and development of the low power consumption technology by the layered confguration of the storage.

(c) High-performance and high reliable interconnection

A design and evaluation of the first version of the 10 gigabit Ethernet switch chip for constituting a large-scale blade server were completed. This switch chip has twelve 10 gigabit Ethernet ports, and, in addition to the high bandwidth, is very suitable for construction of a blade server or a cluster tahnks to the very low latency. Then, design and evaluation of the 10 gigabit Ethernet switch chip of the second version which strengthened the performance to this first version were performed. The enhanced XAUI (eXAUI) macro which can transmit 10Gbps signal over a 25 m copper cable was developed and was included to the second version chip. By this chip, since connection between chassis can be performed without using expensive optical modules, it becomes possible to build a large-scale blade server and a large-scale cluster at reasonable cost. Furthermore, the design and the check of implementability of the third version chip with the transfer function of information of system malfunction were performed.

初年度契約管理番号	５１１０１８０５－０
契約管理番号	０５０００１５６－０